Luo Xi ORCID iD: 0000-0002-0909-9372

Granger Mediation Analysis of Multiple Time Series with an Application to fMRI

Yi Zhao

Department of Biostatistics, Brown University, Providence, Rhode Island, U.S.A.
email: yi_zhao@alumni.brown.edu

and

Xi Luo

Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston, Houston, Texas, U.S.A.
email: xi.rossi.luo@gmail.com

SUMMARY: This paper presents Granger Mediation Analysis (GMA), a new framework for causal mediation analysis of multiple time series. This framework is motivated by a functional magnetic resonance imaging (fMRI) experiment where we are interested in estimating the mediation effects between a randomized stimulus time series and brain activity time series from two brain regions. The independent observation assumption is thus unrealistic for this type of time series data. To address this challenge, our framework integrates two types of models: causal mediation analysis across the mediation variables, and vector autoregressive (VAR) models across the temporal observations. We use "Granger" to refer to VAR correlations modeled in this paper. We further extend this framework to handle multilevel data, in order to model individual variability and correlated errors between the mediator and the outcome variables. Using Rubin's potential outcome framework, we show that the causal mediation effects are identifiable under our time series model. We further develop computationally efficient algorithms to maximize our likelihood-based estimation criteria. Simulation studies show that our method reduces the estimation bias and improves statistical power, compared with existing approaches. On a real fMRI data set, our approach quantifies the causal effects through a brain pathway, while capturing the dynamic dependence between two brain regions.

KEY WORDS: Spatiotemporal dependence; Structural equation modeling; Vector autoregressive models.

This is the author's manuscript of the article published in final edited form as:

Zhao, Y., & Luo, X. (2019). Granger mediation analysis of multiple time series with an application to functional magnetic resonance imaging. Biometrics, 75(3), 788–798. https://doi.org/10.1111/biom.13056

1. Introduction

Mediation analysis is a popular statistical approach for many social and scientific studies. It aims to assess the role of an intermediate variable or mediator sitting in the pathway from a treatment variable to an outcome variable. In many studies, observations from multiple units or subjects are collected, and existing mediation methods usually impose the assumption of independent units explicitly or implicitly. For example, the Baron-Kenny method (Baron and Kenny, 1986; MacKinnon, 2008), built on the structural equation modeling framework, relies on the independence assumption to carry out estimation and inference. Causal mediation analysis has been widely studied in the statistical literature (Imai et al., 2010), and most causal mediation methods again assume independent errors. These methods thus cannot be applied to time series data where temporal dependence is present.

In this paper, we will focus on the time series data generated from a functional Magnetic Resonance Imaging (fMRI) experiment where each participant performs a motor conflict task, responding to randomized STOP/GO experimental stimuli in a sequence of trials. Participants are instructed to press buttons when seeing the GO stimulus, and to withhold from pressing under the STOP stimulus. During the experiment, brain activities are measured by fMRI using the blood-oxygen level dependent (BOLD) contrast. It is well established that the GO stimulus, compared with STOP, will increase brain activation in the primary motor cortex (M1), a brain region responsible for finger movements. Previous studies (Aron et al., 2007; Duann et al., 2009) also discovered that multiple other brain regions also respond to the STOP/GO stimuli. One brain region, the presupplementary motor area (preSMA), was hypothesized to be one of the primary areas for processing the stimuli and mediating the M1 response, though some researchers were not convinced about the primary role of preSMA. Obeso et al. (2013) used repetitive transcranial magnetic stimulation (rTMS) to demonstrate the existence of a brain pathway from preSMA to M1. However, it remained unclear to what extent preSMA mediates the stimulus effect on M1. This question cannot be addressed using widely available neuroimaging analysis tools, because they usually analyze either

the stimulus activations or the connectivity (correlations) between regions. This paper tackles this scientific question by developing a new causal mediation model. In this data example, the stimulus is the treatment variable, and we model the hemodynamic response delay by convolving with a standard hemodynamic response function (Lindquist, 2008). The BOLD activities in preSMA and M1 are the mediator and outcome variables, respectively. All these variables are time series, and an example of these three time series from one participant is shown in Figure 1.

It has been well established before that BOLD time series by fMRI have non-ignorable temporal correlations, which can be effectively modeled by stationary autoregressive (AR) models with a small lag order (Lindquist, 2008). Indeed, autoregressive modeling is an important approach for time series analysis widely studied in the fields of economics and statistics. One earlier approach, named as Granger causality (Granger, 1969, 1980), assesses if the current value of time series x can be predicted by the past values of time series x and another time series y. Such predictive relationship is usually modeled linearly by autoregressive models. This idea is generalized to model multiple time series using multivariate autoregressive models (MAR), also known as vector autoregressive models. Popular estimation methods include (generalized) least squares, the Yule-Walker moments estimator, and maximum likelihood (Lütkepohl, 2005). In particular, Johansen (1991) proposed a conditional maximum likelihood estimator for MAR, using the likelihood of the time series samples in later periods conditional on the time series from the initial periods. Recently, MAR is becoming increasingly popular in fMRI analysis, for example, the implementations in Harrison et al. (2003) and Goebel et al. (2003). Despite its growing popularity, researchers often consider it as a model for "predictive causality", and practitioners need to be careful about the causal interpretation and assumptions (Granger, 2004; Maziarz, 2015). For trivariate time series, conditional Granger causality analysis (Geweke, 1984) is often used to construct test statistics for the "indirect" and "direct" effects. However, these effects are defined differently from the causal mediation effects constructed using potential outcomes. In this paper, we will further develop a multilevel mediation model for time series data, where the temporal correlations are modeled by MAR. To estimate the mediation and MAR parameters jointly in our model, we will further develop the conditional likelihood principle (Johansen, 1991). To the best of our knowledge, causal mediation models of multiple stationary autoregressive

time series have not been studied before, especially when data are multilevel like those collected in our fMRI experiment.

Inferring stimulus effects on BOLD responses has been a central topic in neuroimaging analysis. They are usually implemented using massive linear regressions (Lindquist, 2008). For randomized stimuli, Luo et al. (2012) studied the causal stimulus effects using potential outcomes and nonparametric tests. Sobel and Lindquist (2014) proposed a parametric causal inference framework for fMRI time series, and formulated the causal assumptions using potential outcomes. Some of their assumptions overlap those used by Granger (2004). Recently, several papers used mediation analysis for fMRI to provide further understanding of the causal mechanisms and pathways. Atlas et al. (2010) applied mediation analysis to study the brain mediators of a self-reported behavioral outcome. They utilized a general linear model (GLM) approach to extract the brain activities for each trial (sometimes referred to as single-trial betas), and thus these coefficients in their mediation model can be considered independent. Lindquist (2012) proposed a functional mediation model with fMRI mediators and a scalar outcome. With also a scalar behavioral outcome, Chén et al. (2017) recently proposed multiple mediator models where none of the mediators is modeled as time series. Zhao and Luo (2014) proposed a multilevel causal mediation framework for single-trial betas as the mediator and the outcome. It addresses the issues related to unmeasured confounding and individual variation, but did not directly model the temporal dependence in fMRI time series. Built on the causal framework of Sobel and Lindquist (2014), this paper will extend the multilevel mediation framework to a time series setting. We use "Granger" in this paper to mean temporal MAR correlations, rather than Granger causality.

In a related setting for longitudinal data, marginal structural models (Robins et al., 2000) were employed to quantify causal mediation effects for time-varying treatments and mediators (VanderWeele, 2015). VanderWeele and Tchetgen Tchetgen (2017) introduced the mediational g-formula to estimate the interventional analogs of the natural direct and indirect effects using a semiparametric approach. Lin et al. (2017) later proposed a fully parametric g-formula approach to improve statistical efficiency, especially when the exposure and the mediator are continuous. In these longitudinal mediation models, the outcome of interest is often measured at one time point (at the end) rather than a time series. Their temporal dependence models are also different from our parametric MAR model.

We address these methodological limitations by proposing a new framework, called Granger Mediation Analysis (GMA). It is a mediation model for three MAR time

series. A conceptual diagram of our model is illustrated in Figure 2. Compared with standard mediation models, this model allows the error time series to have more complex dependencies to be discussed later. The causal interpretation of the model parameters will be presented in Section 2.1.

This paper is organized as follows. In Section 2, we introduce our Granger Mediation Analysis framework, which consists of a lower-level mediation model (Section 2.2) and a two-level mediation model (Section 2.4). We compare our method with existing methods through simulation studies in Section B of the supporting information and an analysis of the fMRI data set in Section 3. Section 4 summarizes this paper with discussions and future work.

2. Model and Methods

In this section, we first introduce our single-level GMA model for the time series data from each participant i (Sections 2.1–2.3). To keep the following discussion uncluttered, we drop the participant index i hereafter. In Section 2.4, we will extend this model to multilevel data from multiple participants.

2.1 *Causal definitions*

Our approach builds on Rubin's potential outcome framework (Rubin, 2005). To model fMRI potential responses, we adopt the causal fMRI model and the five causal assumptions (denoted as (SL1)–(SL5) here) in Sobel and Lindquist (2014) (details in Web Section A.1). Readers interested in other applications may skip to Equation (2). Briefly, these five assumptions are: (SL1) BOLD response decomposition; (SL2) true response time invariance; (SL3) temporal consistency; (SL4) p period carry-over; (SL5) no treatment by period interaction. We also assume the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980) that one participant's outcomes do not depend on the treatment assignments of other participants. For $t = 1, \ldots, T$ equally spaced time periods, define $s_{qt} = 1$ if stimulus q is applied at time t and 0 otherwise. In our experiment, we only need to consider two randomized stimuli: $q = 1$ (GO) and $q = 2$ (STOP). Let $\mathbf{s}_t = (s_{1t}, s_{2t})$ be the stimulus assignment at time t, and $\bar{\mathbf{s}}_t = (\mathbf{s}_1, \ldots, \mathbf{s}_t)$ be the historical stimulus assignment up to time t. Following Sobel and Lindquist (2014), we first write the following model for the potential BOLD response of the mediator region

$$\tilde{M}_t(\bar{\mathbf{s}}_t) = \nu_0 + f_{1t}(\bar{\mathbf{s}}_t)a_1 + f_{2t}(\bar{\mathbf{s}}_t)a_2 + \mathbf{N}(\bar{\mathbf{s}}_t)\boldsymbol{\nu} + \mathbf{N}\boldsymbol{\nu}' + \varepsilon_t^{(M)}(\bar{\mathbf{s}}_t), \tag{1}$$

where ν_0 is the intercept, $f_{qt}(\bar{\mathbf{s}}_t) = \sum_{j=0}^{p} s_{q,t-j} h_j$ is the convolution between the stimulus q and the canonical hemodynamic response function h (Lindquist, 2008), $(\mathbf{N}(\bar{\mathbf{s}}_t), \mathbf{N})$ is a vector of measured "nuisance" factors, and $(a_1, a_2, \boldsymbol{\nu}, \boldsymbol{\nu}')$ is a vector of coefficients. The Gaussian error $\varepsilon_t^{(M)}(\bar{\mathbf{s}}_t)$ is assumed to be a zero-mean autoregressive process. This model is essentially the same as the model (model (5)) in Sobel and Lindquist (2014) under a single subject setting. We will discuss the strategy to average the parameter estimates across subjects in Section 2.4.

In fMRI analysis, neuroscientists are often interested in the contrasts between the hemodynamic responses under different stimuli, because BOLD measures have arbitrary units. In our experiment, we are interested in modeling the coefficient contrast $\alpha := a_2 - a_1$, which is interpreted as the effect of the STOP stimulus relative to the GO stimulus. We thus consider a simple "modified covariate" approach below, in the same spirit of the proposal in Tian et al. (2014). Moreover, it is a common practice to preprocess the raw BOLD data by removing the effects of those nuisance covariates, such as head motion and machine drift. Motivated by these two points, we model the "adjusted" mediator response as

$$\begin{aligned} M_t(\bar{\mathbf{s}}_t) &= \tilde{M}_t(\bar{\mathbf{s}}_t) - \gamma_0 - (f_{2t}(\bar{\mathbf{s}}_t) + f_{1t}(\bar{\mathbf{s}}_t))a_1 - \mathbf{N}(\bar{\mathbf{s}}_t)\boldsymbol{\nu} - \mathbf{N}\boldsymbol{\nu}' \\ &= f_{2t}(\bar{\mathbf{s}}_t)(a_2 - a_1) + \varepsilon_t^{(M)}(\bar{\mathbf{s}}_t) \\ &= Z_t(\bar{\mathbf{s}}_t)\alpha + \varepsilon_t^{(M)}(\bar{\mathbf{s}}_t), \end{aligned} \tag{2}$$

where $Z_t(\bar{\mathbf{s}}_t) = f_{2t}(\bar{\mathbf{s}}_t)$. This adjustment also makes the computation later more trackable and the presentation more focused on our mediation model. Using the adjusted brain responses, we propose the following model

$$R_t(\bar{\mathbf{s}}_t, M_t(\bar{\mathbf{s}}_t^*)) = Z_t(\bar{\mathbf{s}}_t)\gamma + M_t(\bar{\mathbf{s}}_t^*)\beta + \varepsilon_t^{(R)}(\bar{\mathbf{s}}_t, \bar{\mathbf{s}}_t^*). \tag{3}$$

Note that this model includes nested counterfactual $R_t(\bar{\mathbf{s}}_t, M_t(\bar{\mathbf{s}}_t^*))$, the potential outcome when the stimulus is set to $\bar{\mathbf{s}}_t$ and M_t to the value when the stimulus is set to $\bar{\mathbf{s}}_t^*$.

Following the standard mediation definitions, we define the average total causal effect by comparing the potential outcomes under the treatment assignment $\bar{\mathbf{s}}_t$ and $\bar{\mathbf{s}}_t^*$ as

$$\mathrm{ATE}(\bar{\mathbf{s}}_t, \bar{\mathbf{s}}_t^*) = \mathbb{E}\left\{R_t(\bar{\mathbf{s}}_t, M_t(\bar{\mathbf{s}}_t)) - R_t(\bar{\mathbf{s}}_t^*, M_t(\bar{\mathbf{s}}_t^*))\right\} = \{Z_t(\bar{\mathbf{s}}_t) - Z_t(\bar{\mathbf{s}}_t^*)\}(\gamma + \alpha\beta).$$

The coefficient $\gamma + \alpha\beta$ is interpreted as the effect on the outcome response for each unit change in $Z_t(\bar{\mathbf{s}}_t) - Z_t(\bar{\mathbf{s}}_t^*)$ due to the stimulus assignment change. A similar formulation for the stimulus effect on a brain region is defined in Sobel and Lindquist (2014).

Under our mediation model, this average total effect is decomposed as the sum of the

average (natural) indirect effect (AIE) and the average (natural) direct effect (ADE)

$$\begin{aligned}\mathrm{ATE}(\bar{\mathbf{s}}_t, \bar{\mathbf{s}}_t^*) &= \mathbb{E}\left\{R_t(\bar{\mathbf{s}}_t, M_t(\bar{\mathbf{s}}_t)) - R_t(\bar{\mathbf{s}}_t, M_t(\bar{\mathbf{s}}_t^*))\right\} + \mathbb{E}\left\{R_t(\bar{\mathbf{s}}_t, M_t(\bar{\mathbf{s}}_t^*)) - R_t(\bar{\mathbf{s}}_t^*, M_t(\bar{\mathbf{s}}_t^*))\right\} \\ &= \mathrm{AIE}(\bar{\mathbf{s}}_t, \bar{\mathbf{s}}_t^*) + \mathrm{ADE}(\bar{\mathbf{s}}_t, \bar{\mathbf{s}}_t^*),\end{aligned}$$

where the two terms above are

$$\mathrm{AIE}(\bar{\mathbf{s}}_t, \bar{\mathbf{s}}_t^*) = \{Z_t(\bar{\mathbf{s}}_t) - Z_t(\bar{\mathbf{s}}_t^*)\}\alpha\beta, \quad \text{and} \quad \mathrm{ADE}(\bar{\mathbf{s}}_t, \bar{\mathbf{s}}_t^*) = \{Z_t(\bar{\mathbf{s}}_t) - Z_t(\bar{\mathbf{s}}_t^*)\}\gamma.$$

The coefficient $\alpha\beta$ represents the effect on the outcome region that is mediated by the mediator region, for each unit change in $Z_t(\bar{\mathbf{s}}_t) - Z_t(\bar{\mathbf{s}}_t^*)$. The coefficient γ represents the effect not mediated by the mediator.

2.2 *A mediation model for time series and causal assumptions*

Because fMRI data are usually preprocessed with various adjustments (not relevant for understanding our mediation method here), throughout the paper we will refer the preprocessed and adjusted fMRI data as the "observed" data. This also makes our method description relevant to other applications when no preprocessing adjustment is required. For the observed data (Z_t, M_t, R_t), $t = 1, \ldots, T$, we first rewrite models (2) and (3) as

$$M_t = Z_t\alpha + E_{1t}, \tag{4}$$

$$R_t = Z_t\gamma + M_t\beta + E_{2t}, \tag{5}$$

where E_{1t} and E_{2t} are two zero-mean error processes. Again, all variables are centered, so no intercepts are included in above. To account for the spatio-temporal dependence between the two error processes, E_{1t} and E_{2t} are assumed to follow a multivariate autoregressive model of order p (MAR(p)):

$$E_{1t} = \sum_{j=1}^{p}\left(\omega_{11_j}E_{1,t-j} + \omega_{21_j}E_{2,t-j}\right) + \epsilon_{1t}, \tag{6}$$

$$E_{2t} = \sum_{j=1}^{p}\left(\omega_{12_j}E_{1,t-j} + \omega_{22_j}E_{2,t-j}\right) + \epsilon_{2t}, \tag{7}$$

where the error vector $(\epsilon_{1t}, \epsilon_{2t})^\top$ is assumed to be a bivariate Gaussian white noise process as

$$\begin{pmatrix}\epsilon_{1t}\\ \epsilon_{2t}\end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix}\sigma_1^2 & \delta\sigma_1\sigma_2\\ \delta\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}. \tag{8}$$

Here $(\epsilon_{1t}, \epsilon_{2t})^\top$ is independent of $(\epsilon_{1u}, \epsilon_{2u})^\top$ for $t \neq u$. Univariate autoregressive errors were considered in Sobel and Lindquist (2014) because they modeled the stimulus effect on each voxel/region separately. For the bivariate errors, we introduce $\mathbf{\Sigma}$ and MAR(p) for the spatio-temporal correlations, where p is usually small (1 or 2) for fMRI data (Lindquist, 2008).

We introduce the correlation parameter δ in (8) to model the instantaneous mediator-outcome dependence, and such dependence (when $\delta \neq 0$) can be due to another unmeasured zero-mean Gaussian process U_t as in the following example. This example is adapted from a sensitivity analysis model for independent observations in Imai et al. (2010). Suppose

$$\epsilon_{it} = g_i U_t + \tilde{\epsilon}_{it}, \qquad i = 1, 2, \tag{9}$$

where $(U_t, \tilde{\epsilon}_{1t}, \tilde{\epsilon}_{2t})$ are mutually independent and also independent of $(U_u, \tilde{\epsilon}_{1u}, \tilde{\epsilon}_{2u})$ for $t \neq u$. It is easy to see that the correlation parameter $\delta \neq 0$ whenever $g_1 g_2 \neq 0$. δ can be interpreted as the magnitude of the unmeasured confounding effect. Figure 2 shows a special case of our proposed model with $p = 1$.

Granger causality analysis in economics (Granger, 1969, 1980) is usually implemented using MAR(p), and recently it has been widely adopted in neuroimaging for so-called Granger connectivity analysis (Harrison et al., 2003; Goebel et al., 2003). We thus name our method as *Granger Mediation Analysis (GMA)*. It is important to note that we use the term "Granger" here to refer to the temporal dependence in the error processes, and we do not intend to interpret these error dependence parameters as causal.

To identify the causal effects (AIE and ADE) from the observed data, we impose the following assumptions.

(A1) The treatment randomization regime is the same across time and participants.

(A2) Models are correctly specified, and there is no treatment-mediator interaction.

(A3) At each time point t, the observed outcome is one realization of the potential outcome with observed treatment assignment $\bar{\mathbf{S}}_t$, where $\bar{\mathbf{S}}_t = (\mathbf{S}_1, \ldots, \mathbf{S}_t)$.

(A4) The treatment assignment is fully random across time, that is,

$$\{R_t(\bar{\mathbf{s}}_t, m_t), M_t(\bar{\mathbf{s}}_t^*)\} \perp\!\!\!\perp \mathbf{S}_t, \ \mathbf{S}_t \perp\!\!\!\perp \mathbf{S}_u \text{ for any } t \neq u, \text{ and } \mathbb{P}(\mathbf{S}_t = \mathbf{s}_t) > 0 \text{ for all } \mathbf{s}_t.$$

(A5) From models (2) and (3), the causal effects are defined based on $(Z_t(\bar{\mathbf{s}}_t), M_t(\bar{\mathbf{s}}_t))$ and $(Z_t(\bar{\mathbf{s}}_t), M_t(\bar{\mathbf{s}}_t^*), R_t(\bar{\mathbf{s}}_t, M_t(\bar{\mathbf{s}}_t^{**}))$ at the same t, and the causal parameters are time-invariant.

(A6) The time-invariant covariance matrix of the Gaussian errors in models (2) and (3)

is not affected by the treatment assignments. That is,

$$\text{Cov}\left[\left\{\varepsilon_t^{(M)}(\bar{\mathbf{s}}_t^*), \varepsilon_t^{(R)}(\bar{\mathbf{s}}_t, \bar{\mathbf{s}}_t^{**})\right\}\right] = \text{Cov}\left[\left\{\varepsilon_t^{(M)}(\bar{\mathbf{s}}_t), \varepsilon_t^{(R)}(\bar{\mathbf{s}}_t, \bar{\mathbf{s}}_t)\right\}\right] = \boldsymbol{\Sigma}$$

for all t, where Cov$[\cdot]$ is the covariance matrix of the vector random variable inside.

Assumptions (A1)–(A3) are adapted from standard causal mediation assumptions (Imai et al., 2010; VanderWeele, 2015). Assumptions (A1) and (A4) are expected to hold in our experiment because the treatment $\mathbf{S}_t$ for every t is randomized and the probability of $\mathbf{S}_t = \mathbf{s}_t$ (for all possible $\mathbf{s}_t$) is the same for all participants and all t in the experiment. Assumption (A4) also satisfies the randomization assumptions for time-varying treatments (Robins and Hernán, 2008), and is expected to hold in our scientific experiment because the stimuli are randomly generated before seeing the fMRI data. Assumptions (A2)–(A3) are regularity conditions for our modeling approach, and these two implicitly assume (SL1)–(SL5) from Sobel and Lindquist (2014). Assumption (A2) implies the parametric assumptions, such as linearity and Gaussian errors, in our model. Assumption (A3) is also known as the "consistency" assumption in causal inference (VanderWeele, 2009). Assumption (A5) considers only the effects at each time point in this paper, because fMRI has low temporal resolution and many fMRI analysis methods study only the effects between regions at the same time point t (though the actual measurement times between the two regions may differ by an amount smaller than the sampling frequency). This assumption is similar in spirit to the "short-term" effect considered in Keogh et al. (2017). In Assumption (A6), we replace the ignorability assumptions of the mediator (Imai et al., 2010; VanderWeele and Tchetgen Tchetgen, 2017) by a Gaussian covariance assumption. Because all errors are multivariate Gaussian (regardless of treatment assignments), setting $\delta = 0$ implies the so-called "cross-world" independence assumption. δ can also be treated as a sensitivity parameter as in Imai et al. (2010). When multilevel data are available as in our experiment, it can also be fitted using another second level parametric model (Zhao and Luo, 2014) across participants with additional assumptions to be discussed later. This approach essentially assumes that the constant effect of unmeasured mediator-outcome confounder U_t is fully characterized by the error correlation δ, as discussed in the example (9) before. The correlation parameter δ is also assumed to be constant across time, and this is similar to the constant causal effect assumption in Granger (1980).

For the MAR(p) models (6) and (7), we impose the following stationary condition for parameter estimation.

(A7) The eigenvalues of the companion matrix have modulus less than one.

The companion matrix is given in Section A.3 of the supporting information. Assumption (A7) is a standard condition for stationary autoregressive processes (Lütkepohl, 2005). This stationarity condition is deemed satisfied for adjusted fMRI data after correcting for the stimulus effects and other covariates (Harrison et al., 2003; Chang and Glover, 2010), as in our model.

2.3 *Method*

In this section, we extend the maximum (conditional) likelihood estimation for MAR (Johansen, 1991) to our Granger mediation model. To derive the likelihood, we note the following equivalent formulation for models (4)–(7) of the observed data:

$$M_t = Z_t\alpha + \sum_{j=1}^{p}\left(\phi_{1j}Z_{t-j} + \psi_{11_j}M_{t-j} + \psi_{21_j}R_{t-j}\right) + \epsilon_{1t}, \tag{10}$$

$$R_t = Z_t\gamma + M_t\beta + \sum_{j=1}^{p}\left(\phi_{2j}Z_{t-j} + \psi_{12_j}M_{t-j} + \psi_{22_j}R_{t-j}\right) + \epsilon_{2t}, \tag{11}$$

where $\{\phi_{1j}, \phi_{2j}, \psi_{11_j}, \psi_{21_j}, \psi_{12_j}, \psi_{22_j}\}$ are the new parameters introduced to facilitate our likelihood formulation, and we don't intend to interpret them individually. The variance parameters for $(\epsilon_{1t}, \epsilon_{2t})$ are $(\sigma_1, \sigma_2, \delta)$ given in (8). To see the equivalence, one can plug (6) and (7) into (4) and (5), respectively, and then replace respectively $E_{1,t-j}$ and $E_{2,t-j}$ by $M_{t-j} - Z_{t-j}\alpha$ and $R_{t-j} - Z_{t-j}\gamma - M_{t-j}\beta$, for $j = 1, \ldots, p$.

The parameters in these two equivalent formulations have an explicit linear relationship shown by Web Lemma A.1. We thus propose to estimate the parameters in models (4)–(7) by transforming the parameter estimates obtained from models (10)–(11).

Our formulation (10)–(11) is a linear structural equation model with correlated errors between two equations. Therefore, one cannot fit (10) and (11) separately, using for example standard (generalized) least squares for autoregressive models. We propose an estimation approach based on the principle of maximizing the conditional likelihood.

To simplify the notation in our derivation, we introduce the following matrix representations: let $\boldsymbol{\theta}_1 = (\alpha, \boldsymbol{\phi}_1^\top, \boldsymbol{\psi}_{11}^\top, \boldsymbol{\psi}_{21}^\top)^\top$, $\boldsymbol{\theta}_2 = (\gamma, \boldsymbol{\phi}_2^\top, \boldsymbol{\psi}_{12}^\top, \boldsymbol{\psi}_{22}^\top)^\top$, where $\boldsymbol{\phi}_j = (\phi_{j_1}, \ldots, \phi_{j_p})^\top$, $\boldsymbol{\psi}_{jk} = (\psi_{jk_1}, \ldots, \psi_{jk_p})^\top$ for $j, k = 1, 2$; $\mathbf{X}_t = (Z_t, \mathbf{Z}_{t-1}^{(p)\top}, \mathbf{M}_{t-1}^{(p)\top}, \mathbf{R}_{t-1}^{(p)\top})^\top$, where $\mathbf{Z}_{t-1}^{(p)} = (Z_{t-1}, \ldots, Z_{t-p})^\top$, and $\mathbf{M}_{t-1}^{(p)}$ and $\mathbf{R}_{t-1}^{(p)}$ are defined analogously. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \beta, \sigma_1, \sigma_2)$ be all the model parameters except δ. Given the initial p time periods, the conditional

log-likelihood (ignoring constants) is

$$\begin{aligned}\ell\left(\boldsymbol{\Theta}, \delta \mid \mathbf{Z}, \mathcal{I}_p\right) &= \sum_{t=p+1}^{T} \log f\left((M_t, R_t) \mid \mathbf{X}_t\right) \\ &= -\frac{T-p}{2}\log \sigma_1^2\sigma_2^2(1-\delta^2) - \frac{1}{2\sigma_1^2}\|\mathbf{M}-\mathbf{X}\boldsymbol{\theta}_1\|_2^2 \\ &\quad -\frac{1}{2\sigma_2^2(1-\delta^2)}\|(\mathbf{R}-\mathbf{M}\beta-\mathbf{X}\boldsymbol{\theta}_2)-\kappa(\mathbf{M}-\mathbf{X}\boldsymbol{\theta}_1)\|_2^2, \qquad (12)\end{aligned}$$

where $\mathcal{I}_p = \{(Z_1, M_1, R_1), \ldots, (Z_p, M_p, R_p)\}$ is the initial p observations; f is the likelihood for (M_t, R_t) conditioning on the previous p periods; $\|\mathbf{x}\|_2$ is the ℓ_2-norm of vector $\mathbf{x}$; $\mathbf{R} = (R_{p+1}, \ldots, R_T)^\top$, similarly for $\mathbf{M}$ and $\mathbf{X}$; T is the number of time points; and $\kappa = \delta\sigma_2/\sigma_1$.

In our model, δ accounts for the effect of an unmeasured confounding process to the mediator and the outcome, for example, model (9) in Section 2.2. In the classical mediation analysis setting when data are collected from independent units, a similarly defined parameter δ is not identifiable from observed data, and thus it is often treated as a sensitivity parameter to account for the effect of unmeasured confounding (Imai et al., 2010).

Though we cannot estimate δ from the conditional likelihood of single-level data (Web Theorem A.2), we show that our estimators for β and γ are expressed as functions of δ. The conditional maximum likelihood estimator (CMLE) of all the remaining parameters is given in explicit forms in Web Section A.5. In there, we also show that our estimators for β and γ are consistent after correcting for δ, and the asymptotic covariance matrix is derived in Web Theorem A.4. Based on these results, our method allows δ to be treated as a parameter in sensitivity analysis. We illustrate these points using a toy simulation example in Web Section B.1. In Section 2.4, we consider an alternative approach to estimate δ by maximizing a second level likelihood function of all the estimates of α, β, γ pooled across participants.

2.4 *Extension to two-level data*

In this section, we extend our GMA model for the two-level time series data in our fMRI experiment, adapting the multilevel mediation method for independent observations proposed by Zhao and Luo (2014).

2.4.1 *Model.* We will refer to the two levels as participant and scan time in this paper. For the time series of participant i ($i = 1, \ldots, N$), we model the first-level scan-time data by our single level GMA model (4)–(7), and all the modeling parameters should now

be denoted with subscript i. For example, α_i, β_i and γ_i are the causal parameters of participant i. In order to estimate the population averages of the causal effects and account for the between-participant variations, we employ the following multivariate linear model

$$\boldsymbol{\vartheta}_i = \boldsymbol{\vartheta} + \boldsymbol{\eta}_i, \tag{13}$$

where $\boldsymbol{\vartheta}_i = (\alpha_i, \beta_i, \gamma_i)^\top$; $\boldsymbol{\vartheta} = (\alpha, \beta, \gamma)^\top$ denotes the population level coefficients; and $\boldsymbol{\eta}_i = (\epsilon_i^\alpha, \epsilon_i^\beta, \epsilon_i^\gamma)^\top$ is the random error of participant i, which is assumed to be independent and identically distributed from a trivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Lambda}$. The linear additive form in (13) for modeling the population and individual parameters is standard in fMRI analysis (Lindquist, 2008). At the population level, the population direct effect is γ, and the population indirect effect is $\alpha\beta$ by the product method. There is an alternative definition of the population indirect effect by the difference method (Kenny et al., 2003). This approach would also require fitting a total effect model by regressing outcome R on treatment Z for each participant, and a population equation analogous to model (13). For the sake of space, we omit the description of this alternative approach in this paper, because they yield very similar numerical results.

As discussed in Section 2.3, we estimate the parameters through the equivalent formulation using $\mathbf{X}_{i_t}$, $\boldsymbol{\theta}_{i_1}$ and $\boldsymbol{\theta}_{i_2}$ defined in the same way as in Section 2.3 for participant i, $i = 1, \ldots, N$. Let δ_i be the error correlation between $\epsilon_{i_{1t}}$ and $\epsilon_{i_{2t}}$. As shown in Web Theorem A.2, δ_i is not estimable from the individual conditional likelihood function for each participant i. Because the joint likelihood of N independent participants is simply a product of individual likelihood functions, one cannot estimate different δ_i from the joint likelihood function either. In order to estimate δ_i from data, we adopt the optimization methods in Zhao and Luo (2014), and impose the following assumption.

(A8) δ_i is constant across participants, i.e., $\delta_i = \delta$ for all i.

Without assumption (A8), one may propose to perform sensitivity analysis using different δ_i for each i. However, the number of sensitivity parameters makes this proposal computationally unrealistic for large N. Assumption (A8) reduces the number of parameters in our model and allows our model to pool information across subjects to estimate a single δ. We will introduce two algorithms to estimate δ in the next section.

2.4.2 *Method.* The principal idea in Zhao and Luo (2014) is to estimate δ by maximizing the joint likelihood of N participants. We adopt this idea for our GMA model

here. Let $\boldsymbol{\Upsilon} = (\delta, \boldsymbol{\vartheta}, \boldsymbol{\Lambda}, (\boldsymbol{\theta}_{i_1}, \boldsymbol{\theta}_{i_2}, \beta_i), (\sigma_{1_i}, \sigma_{2_i}))$, the conditional log-likelihood function (conditioning on the initial p time points of each subject's data) is written as

$$\begin{aligned} h(\boldsymbol{\Upsilon}) &= \sum_{i=1}^{N} \sum_{t=p+1}^{T_i} \log \mathbb{P}\left(R_{i_t}, M_{i_t} \mid \mathbf{X}_{i_t}, \boldsymbol{\theta}_{i_1}, \boldsymbol{\theta}_{i_2}, \beta_i, \delta, \sigma_{i_1}, \sigma_{i_2}\right) + \sum_{i=1}^{N} \log \mathbb{P}\left(\boldsymbol{\vartheta}_i \mid \boldsymbol{\vartheta}, \boldsymbol{\Lambda}\right) \\ &= h_1 + h_2, \end{aligned} \tag{14}$$

where $\boldsymbol{\vartheta}_i = (\alpha_i, \beta_i, \gamma_i)$, α_i and γ_i are the first element of $\boldsymbol{\theta}_{i_1}$ and $\boldsymbol{\theta}_{i_2}$, respectively; T_i is the number of time points of subject i; h_1 is the sum of N log-likelihood functions (12), and h_2 is the log-likelihood function of model (13). It is challenging to optimize these many parameters that grow with N. In particular, our GMA model also contains several temporal correlation parameters for each subject. We thus develop two algorithms for maximizing the joint likelihood, with different computational complexity and numerical accuracy.

A two-stage algorithm. This algorithm is inspired by the two-level massive linear regression method commonly applied for fMRI analysis, for example in Kenny et al. (2003) and Lindquist (2008). In the first stage, we estimate, for each participant i, the coefficients in the single level model with a given δ using Web Proposition A.3. This stage splits the computation cost by maximizing the summands in h_1 for each participant, which can be computed in parallel. In the second stage, we plug in the estimated coefficients from the first stage into the left-hand side of the second level regression model (13), and we maximize its likelihood function h_2. To estimate δ, we repeat the two-stage computation for different δ, and then use a one-dimensional optimization algorithm (e.g., Newton's method) to find the δ that yields the maximum joint likelihood h.

The key challenge for proving the asymptotic consistency of this algorithm is to show that δ is estimable and consistently estimated using the above algorithm. The consistency of the remaining parameters (given δ) is guaranteed by the standard maximum likelihood theory under regularity conditions.

THEOREM 1: *Assume assumptions (A1)-(A8) are satisfied. Assume* $\mathbb{E}(Z_{i_t}^2) = q < \infty$, *for* $i = 1, \ldots, N$. *Let* $T = \min_i T_i$.

(1) If $\boldsymbol{\Lambda}$ *is known, then the two-stage estimator* $\hat{\delta}$ *maximizes the profile likelihood of model* (13) *asymptotically, and* $\hat{\delta}$ *is* $\sqrt{NT}$*-consistent.*

(2) If $\boldsymbol{\Lambda}$ *is unknown, then the profile likelihood of model* (13) *has a unique maximizer* $\hat{\delta}$ *asymptotically, and* $\hat{\delta}$ *is* $\sqrt{NT}$*-consistent, provided that* $1/\varpi = \bar{\kappa}^2/\varrho^2 = \mathcal{O}_p(1/\sqrt{NT})$, $\kappa_i = \sigma_{i_2}/\sigma_{i_1}$, $\bar{\kappa} = (1/N)\sum \kappa_i$, *and* $\varrho^2 = (1/N)\sum(\kappa_i - \bar{\kappa})^2$.

Using the two-stage estimator $\hat{\delta}$, the CMLE of our model (Web Proposition A.3) is consistent, as well as the estimator for $\boldsymbol{\vartheta} = (\alpha, \beta, \gamma)$ in model (13).

To verify the estimability of δ in practice, we plot the maximum log-likelihood value against δ. Web Figure B.2a illustrates such a plot for a toy simulated data set. The joint likelihood h is unimodal, while the single level likelihood in Web Figure B.1a is flat.

A block coordinate-descent algorithm. Though the two-stage algorithm is computationally fast and asymptotically consistent, it only approximately maximizes the joint likelihood h. To improve the finite sample performance, we propose a block coordinate-descent algorithm for maximizing h_1 and h_2 jointly. Some finite sample improvement was observed by a similar strategy in Zhao and Luo (2014). We propose the following optimization problem

$$\max_{\boldsymbol{\Upsilon}:\big((\sigma_{i_1},\sigma_{i_2}),\boldsymbol{\Lambda}\big)\in\mathcal{S}} h(\boldsymbol{\Upsilon}), \tag{15}$$

where $\mathcal{S}$ is a constraint set for the variance components. We put a positive constraint on $(\sigma_{i_1}, \sigma_{i_2})$, and a positive definite constraint on $\boldsymbol{\Lambda}$. We propose to optimize blocks of variables (except δ) iteratively because the updates for each block of variables are given in explicit forms, conditional on all other variables (Web Theorem A.6). After obtaining the profile likelihood value for each δ, we estimate δ by a one-dimensional optimization algorithm as before. The full algorithm is summarized in Web Algorithm A.1. For this block coordinate-descent algorithm, we also propose to check the solution of δ graphically as before (Web Figure B.2b).

2.5 *Inference*

Because the distribution of the product $\hat{\alpha}\hat{\beta}$ can be far from Gaussian, we propose to employ bootstrap over participants to perform statistical inference on the population causal effects.

3. The fMRI Experiment

The data set was obtained from the OpenfMRI database, and the accession number is ds000030. In the experiment, $N = 121$ right-handed participants in healthy condition were recruited. The participants were asked to perform motor responses to two types of randomized stimuli: GO or STOP. The STOP/GO stimuli were randomly intermixed with 96 GO and 32 STOP stimuli, with randomly jittered time intervals between the stimuli. Under the GO stimulus, the participants should respond with button presses; under the STOP stimulus, the participants should withhold from pressing when a stop

signal (a 500 Hz tone) was presented after the GO stimulus. More details about this experiment can be found in Poldrack et al. (2016). Data preprocessing steps are described in Web Section C.1.

We compare the mediation effect estimates from the proposed block coordinate-descent (GMA-h) and two-stage (GMA-ts) methods with the two-level method in Zhao and Luo (2014) (MACC-h), the multilevel SEM method proposed by Kenny et al. (2003) (KKB) and the Baron-Kenny (BK) method (Baron and Kenny, 1986). Because other competing methods do not provide estimates of the transition matrix, we compare the transition matrix estimates with the MAR fits by Harrison et al. (2003), which does not model the mediation effects. We set the lag parameter $p = 2$ in our GMA approach. We also tried $p = 3$, but the lag-three temporal correlation estimates are close to zero (Web Section C.4). All methods use 200 bootstrap samples for inference.

Table 1 presents the estimates (and the 95% bootstrap confidence intervals) of δ, γ and $\alpha\beta$. The estimates from GMA-ts and GMA-h are close, consistent with our simulations. Specifically, the GMA-h estimates are $\hat{\gamma} = -1.729$ and $\widehat{\alpha\beta} = -0.623$. The negative estimates suggest that the STOP stimulus deactivates M1 both directly and indirectly through the preSMA pathway. The average indirect effect through preSMA is about half of the average direct effect or one-third of the average total effect. This confirms that the mediation effect of preSMA is at least medium, while there may be other pathways that account for a substantial portion of the total effect. Thus future research is needed to explore and understand these other pathways. The estimates of δ by both GMA-ts and GMA-h are negative and significantly different from zero. The nonzero estimates provide evidence of the existence of unmeasured confounding in the data. These two estimates of δ are also close on this dataset. MACC-h produces a larger estimate of δ, which is consistent with the simulation results.

Our GMA-ts and GMA-h estimates of γ and $\alpha\beta$ are different from all other methods. In particular, our GMA methods produce the largest indirect effect estimates in magnitude. MACC-h yields a much smaller estimate (about 30% less) in magnitude. KKB and BK also yield smaller estimates, because they fail to account for the confounding effect or nonzero δ. Though all these estimates give the same qualitative interpretation for the role of preSMA, our quantitative estimates here suggest a much larger role of preSMA than other methods.

Another advantage of our GMA methods is that it also estimates the temporal dependen-

cies between two brain regions, which are represented by the transition matrix $\mathbf{\Omega}$. The estimate of $\mathbf{\Omega}$ is shown in Web Table C.1, where we observe significant feedback effects from M1 to preSMA at lag one and lag two ($\hat{\omega}_{21_1} = 0.100$ and $\hat{\omega}_{21_2} = -0.076$). Comparing with the estimates by MAR (Web Table C.1), we find that MAR produces larger point estimates of the diagonals and has larger variability overall, probably because it does not model the direct and indirect effects like ours. MAR also yields wider confidence intervals for the off-diagonals than ours, though the point estimates are similar. Web Section C.2 presents the impulse response function plots of the MAR models.

4. Discussion

In this paper, we propose a mediation analysis framework for time series data. Our approach integrates multivariate autoregressive models and mediation analysis to yield a better understanding of such data. Our approach is also embedded in a causal mediation model for correlated errors. We prove that a simple two-stage algorithm will yield asymptotically unique and consistent estimates, and its finite sample performance is improved by a more sophisticated optimization algorithm with increased computational costs. Using both simulations and a real fMRI data set, we demonstrate the numerical advantages of our proposal.

Our model setup is motivated by several important statistical models for task-related fMRI data. It is likely that other scientific experiments or studies will require different modeling components, due to different data structures for the treatment, mediator, and outcome. For example, Kenny et al. (2003) discussed various multilevel data sets, where the variables are scalars instead of time series at the participant level. Time series modeling is also a topic with a long history, and some other time series models, other than MAR, may be more suitable for certain experiments. We will explore these different settings in future research. In this paper, we focus on randomized treatment. It is also interesting to further develop our proposal using the tools for observational studies to relax the randomization requirement.

Many extensions of mediation models have also been considered in the literature (VanderWeele, 2015). These models can also include interactions and covariates, which are common in many social studies. Our method relies on the fully parametric assumptions, and our simulation shows that deviating from these assumptions may introduce biases, for example when nonlinear effects are present. We are interested in extending our proposal to these more complex settings in the future.

Acknowledgements

We thank the editor, associate editor, and three anonymous reviewers for their very helpful comments. This work was partially supported by National Institutes of Health grants R01EB022911, P20GM103645, P01AA019072, R01MH110449, and S10OD016366, National Science Foundation grant DMS 1557467.

References

Aron, A. R., Behrens, T. E., Smith, S., Frank, M. J., and Poldrack, R. A. (2007). Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (mri) and functional mri. *Journal of Neuroscience* **27,** 3743–3752.

Atlas, L. Y., Bolger, N., Lindquist, M. A., and Wager, T. D. (2010). Brain mediators of predictive cue effects on perceived pain. *The Journal of neuroscience* **30,** 12964–12977.

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* **51,** 1173.

Chang, C. and Glover, G. H. (2010). Time–frequency dynamics of resting-state brain connectivity measured with fmri. *Neuroimage* **50,** 81–98.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., and Lindquist, M. A. (2017). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19,** 121–136.

Duann, J.-R., Ide, J. S., Luo, X., and Li, C.-s. R. (2009). Functional connectivity delineates distinct roles of the inferior frontal cortex and presupplementary motor area in stop signal inhibition. *The Journal of Neuroscience* **29,** 10171–10179.

Geweke, J. F. (1984). Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association* **79,** 907–915.

Goebel, R., Roebroeck, A., Kim, D.-S., and Formisano, E. (2003). Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magnetic resonance imaging* **21,** 1251–1261.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* pages 424–438.

Granger, C. W. (1980). Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control* **2,** 329–352.

Granger, C. W. (2004). Time series analysis, cointegration, and applications. *American Economic Review* **94,** 421–425.

Harrison, L., Penny, W. D., and Friston, K. (2003). Multivariate autoregressive modeling of fmri time series. *NeuroImage* **19,** 1477–1491.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* pages 51–71.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: Journal of the Econometric Society* pages 1551–1580.

Kenny, D. A., Korchmaros, J. D., and Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological methods* **8,** 115.

Keogh, R. H., Daniel, R. M., Vanderweele, T. J., and Vansteelandt, S. (2017). Analysis of longitudinal studies with repeated outcome measures: adjusting for time-dependent confounding using conventional methods. *American journal of epidemiology* **187,** 1085–1092.

Lin, S.-H., Young, J., Logan, R., Tchetgen, E. J. T., and VanderWeele, T. J. (2017). Parametric mediational g-formula approach to mediation analysis with time-varying exposures, mediators, and confounders. *Epidemiology* **28,** 266–274.

Lindquist, M. A. (2008). The statistical analysis of fmri data. *Statistical Science* **23,** 439–464.

Lindquist, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association* **107,** 1297–1309.

Luo, X., Small, D., Li, C., and Rosenbaum, P. (2012). Inference with interference between units in an fmri experiment of motor inhibition. *Journal of the American Statistical Association* **107,** 530–541.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Routledge.

Maziarz, M. (2015). A review of the granger-causality fallacy. *The journal of philosophical economics: Reflections on economic and social issues* **8,** 86–105.

Obeso, I., Cho, S. S., Antonelli, F., Houle, S., Jahanshahi, M., Ko, J. H., and Strafella, A. P. (2013). Stimulation of the pre-sma influences cerebral blood flow in frontal areas involved with inhibitory control of action. *Brain stimulation* **6,** 769–776.

Poldrack, R., Congdon, E., Triplett, W., Gorgolewski, K., Karlsgodt, K., Mumford, J., Sabb, F., Freimer, N., London, E., and Cannon, T. (2016). A phenome-wide examination of neural and cognitive function. *Scientific data* **3,** 160110.

Robins, J. M. and Hernán, M. A. (2008). Estimation of the causal effects of time-varying

exposures. In *Longitudinal data analysis*, pages 547–593. Chapman and Hall/CRC.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* pages 550–560.

Rubin, D. B. (1980). Discussion of "randomized analysis of experimental data: the fisher randomization test" by basu d. *Journal of the American Statistical Association* **75,** 591–593.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association* **100,**.

Sobel, M. E. and Lindquist, M. A. (2014). Causal inference for fmri time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *Journal of the American Statistical Association* **109,** 967–976.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association* **109,** 1517–1532.

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology* **20,** 880–883.

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79,** 917–938.

Zhao, Y. and Luo, X. (2014). Estimating mediation effects under correlated errors with an application to fmri. *arXiv preprint arXiv:1410.7217* .

Supporting Information

Additional supporting information may be found online in the Supporting Information Section at the end of the article. An R package implementation of our method is publicly available on CRAN at `https://CRAN.R-project.org/package=gma`.

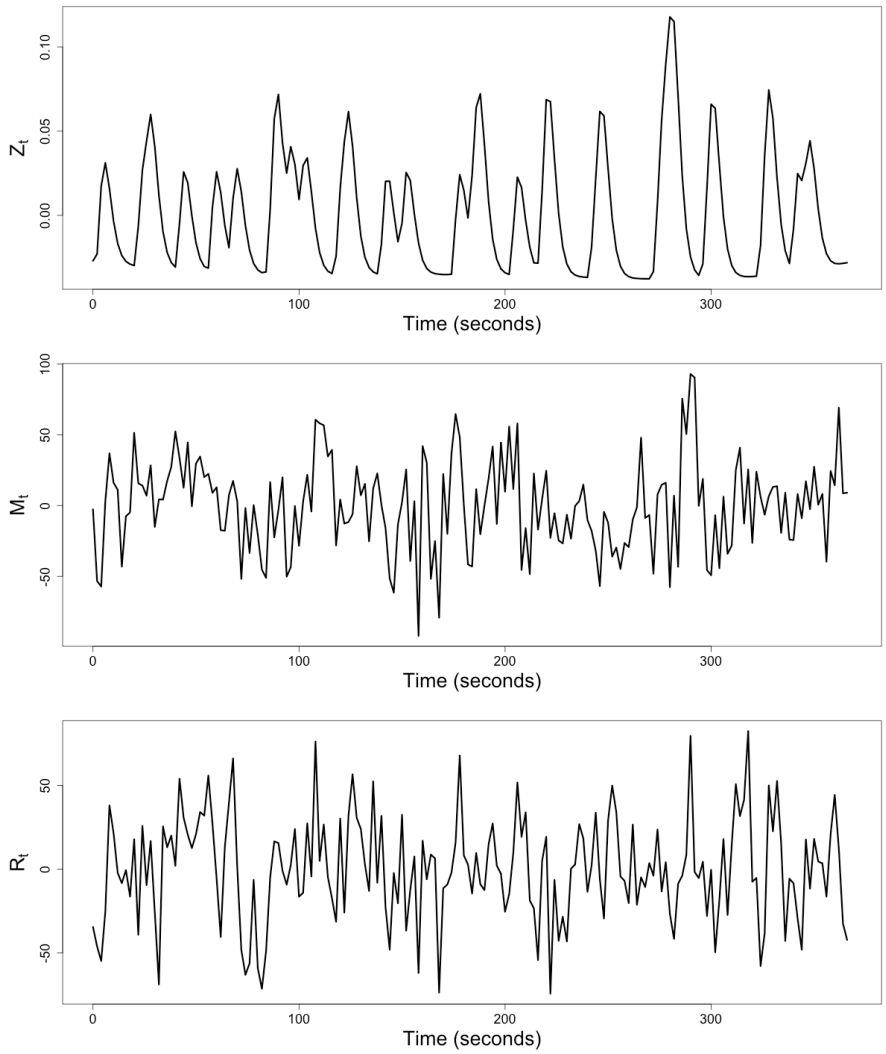

Figure 1: The stimulus input time series ($Z_t = f_t(\bar{\mathbf{S}}_t)$, the convolution of the stimulus $\bar{\mathbf{S}}_t = (\mathbf{S}_1, \ldots, \mathbf{S}_t)$ with the canonical hemodynamic response function), and the preSMA (M_t) and M1 (R_t) fMRI BOLD time series from one of the 121 participants. The notations are given in Section 2.1.

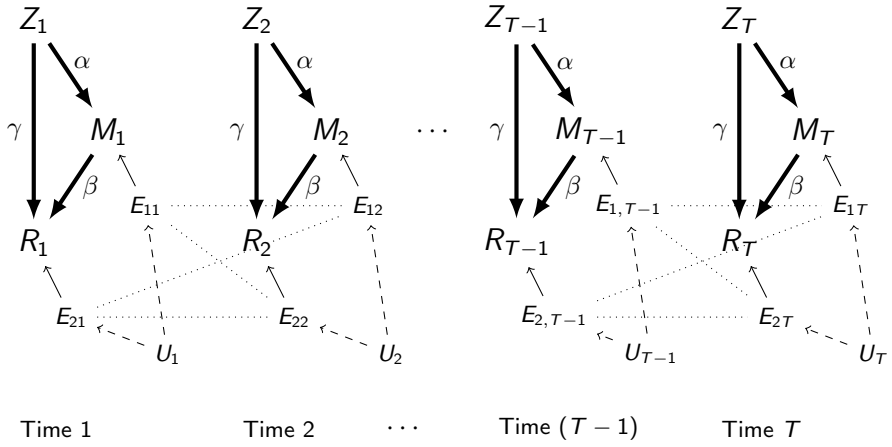

Figure 2: A conceptual diagram for the time series data of a single participant. At each time point, the bold arrows between the convolved stimulus time series Z_t, the mediator M_t and the outcome R_t depict the causal mediation mechanism we aim to study. E_{1t} and E_{2t} are random autoregressive errors. An unmeasured confounding variable U_t influences both errors. Dotted lines represent the autoregressive dependence in our model.

Table 1: Average estimates and 95% confidence intervals from GMA-h, GMA-δ, MACC-h, KKB and BK for the fMRI data set using 200 bootstrap samples. GMA: Granger mediation analysis method proposed in this paper, using either the hierarchical likelihood algorithm (GMA-h) or the two-stage algorithm (GMA-ts); MACC-h: mediation analysis with correlated errors by Zhao and Luo (2014); KKB: multilevel mediation by Kenny et al. (2003); BK: (single-level) mediation by Baron and Kenny (1986).

Method	δ	γ	α	β	$\alpha\beta_p$
GMA-h	-0.370 (-0.530, -0.156)	-1.729 (-2.445, -0.964)	-0.739 (-1.487, 0.035)	0.838 (0.644, 0.999)	-0.623 (-1.239, 0.033)
GMA-ts	-0.343 (-0.501, -0.163)	-1.722 (-2.461, -0.904)	-0.740 (-1.442, 0.080)	0.810 (0.656, 0.965)	-0.604 (-1.234, 0.055)
MACC-h	-0.762 (-0.799, -0.721)	-2.310 (-2.958, -1.641)	-0.259 (-0.810, 0.303)	1.511 (1.410, 1.619)	-0.391 (-1.271, 0.465)
KKB	-	-2.513 (-2.922, -2.073)	-0.225 (-0.772, 0.326)	0.617 (0.589, 0.647)	-0.140 (-0.467, 0.196)
BK	-	-2.583 (-3.023, -2.142)	-0.235 (-0.774, 0.352)	0.616 (0.588, 0.647)	-0.146 (-0.467, 0.211)